# INCREMENTAL ADJUSTMENT OF STATE-DEPENDENT BIAS PARAMETERS FOR ADAPTIVE SPEECH RECOGNITION

## FIELD OF INVENTION

[0001]    This invention relates to speech recognition and more particularly to speech recognition in adverse conditions.

## BACKGROUND OF INVENTION

[0002]    In speech recognition, inevitably the speech recognizer has to deal with recording channel distortions, background noises, and speaker variabilities. The factors can be modeled as mismatch between the distributions of acoustics models (HMMs) and speech feature vectors. To reduce the mismatch, speech models can be compensated by modifying the acoustic model parameters according to the amount of observations collected in the target environment from the target speaker. See Yifan Gong, "Speech Recognition in Noisy Environments ": A survey, Speech Communication, 16(3):pp261-291, April 1995.

[0003]    Currently, in typical recognition systems, batch parameter estimations are employed to update parameter after observation of all adaptation data. See L.A. Liporace, Maximum likelihood estimation for multivariate observations of Markov sources, IEEE Transactions on Information Theory, IT-28(5): pp729-734, September 1982 and L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE,77(2):pp257-285, February 1989. Batch processing can not track parameter variations and is therefore not suitable to follow slow time-varying environments and speaker changes. To deal with noisy background, noise statistics can be collected and used to compensate Model mean vectors. See M.J.F. Gales, PMC for speech recognition in additive and convolutional noise, Technical Report

TR-154, CUED/F-INFENG, December 1993. However it is necessary to obtain an estimate of noises, which in practice is not straight forward since the noise itself may be time varying. Speaker adaptation based on MLLR improves recognition performance. See C.J. Leggetter and P.C. Woodland, Flexible speaker adaptation for large vocabulary speech recognition, IN Proceedings of European Conference on Speech Communication and Technology, Volume II, pages 1155-1158, Madrid Spain, Sept. 1955. It requires, however, that all the adaptation utterances be collected in advance. Sequential parameter estimation has been used for estimating time-varying noises in advance. See K. Yao, K.K. Paliwal, and S. Nakamura, Noise adaptive speech recognition in time-varying noise based on sequential kullback proximal algorithm, In Proc. of Inter. Conf. on Acoustics, Speech and Signal Processing, volume 1, pages 189-192, 2002. However, such formulation does not adapt the system to the speaker and channel.

## SUMMARY OF INVENTION

[0004]    In accordance with one embodiment of the present invention a method of updating bias of a signal model in a sequential manner is provided by introducing an adjustable bias in the distribution parameter of the signals; updating the bias every time a new observation of the signal is available; and calculating the updated new bias by adding a correction item to the old bias.

[0005]    In accordance with another embodiment of the present invention state-dependent bias vectors are added to the mean vectors and adjust them to match a given operation condition. The adjustment is based on the utterances recognized in the past, and no additional data collection is necessary.

[0006]     In accordance with an embodiment of the present invention adapt bias

vector parameters which can be shared , one for each Gaussian, after observing each

utterance ( rather than waiting for all utterances to be available) and scan only once each

utterance (single pass).

DESCRIPTION OF DRAWING

[0007]     Figure 1 illustrates a speech recognizer according to the prior art with

observing and storing N utterances and then update.

[0008]     Figure 2 illustrates Gaussian distributions by plot of amplitude in the Y

axis and frequency in the x axis.

[0009]     Figure 3 illustrated the method according to one embodiment of the

present invention to modify the mean vectors.

[0010]     Figure 4 illustrates all of the states in different frames tied to the same

bias.


DESCRIPTION OF PREFERRED EMBODIMENT OF THE PRESENT INVENTION

[0011]     A speech recognizer as illustrated in Figure 1 includes speech models 13

and speech recognition is achieved by comparing the incoming speech to the speech

models such as Hidden Markov Models (HMMs) models at the recognizer 11. This

invention is about an improved model used for speech recognition. In the traditional

model the distribution of the signal is modeled by a Gaussian distribution defined by $\mu$

and $\Sigma$ where $\mu$ is the mean and $\Sigma$ is the variance. The observed signal $O_t$ is defined by

observation N $(\mu,\Sigma)$. Curve A of Figure 2 illustrates a Gaussian distribution. If you have

noise or any distortion such as a difference speaker or microphone channel the values

change such as represented by curve B of Figure 2. In the prior art Expectation

3

Maximization (EM) approach the procedure is to observe the utterance N and then do an update. The formulation required a specified number of utterances is used to get a good mean bias. There is a need to collect adaptation data and noise statistics. That number may be 1000 with many speakers. This does not permit one to correct for the individuality of the speaker or account for channel changes.

[0012]    The present invention provides sequential bias adaptation (SBA) introduces a bias vector to each of the mean vectors of Gaussian distributions of the recognizer 31 as shown in Figure 3.   It adapts the biases of the acoustic models online sequentially based on the sequential Expectation-Maximization (EM) algorithm. The bias vectors are updated on new speech observations, which may be the utterance just presented to the recognizer 31. The new speech observation may be for every sentence, every word, number dialed, or sensing a quiet and then updating. This permits correcting for the individuality of the speaker and for correcting for channel changes. For sequential bias adaptation, there is no need to explicitly collect adaptation data, and no need to collect noise statistics. The new observation is used with the old bias to calculate the new bias adjustment as illustrated by block 35 and that is used to provide the updated bias adjustment to the models 33.

[0013]    The following equation (1) is the performance index or Q function. The Q function is a function of θ which includes this bias.

$$Q_{K+1}^{(s)}(\Theta_k, \theta) = \sum_{r=1}^{K+1} Q_r(\Theta_k, \theta) \tag{1}$$

where $Q_{k+1}^{(s)}$ denotes the EM auxiliary Q-function based on all the utterances from 1 to k+1, in which  is the parameter set at utterance $k$ and $\theta$ denotes a new parameter set. See

A.P. Dempster, N. M. Laird, and D.B. Rubin "Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(1):1-38, 1977.

[0014]    $Q_{k+1}^{(s)}$ can be written in a recursive way as:

$$Q_{k+1}^{(s)}(\Theta_k,\theta) = Q_k^{(s)}(\Theta_{k-1},\theta) + L_{k+1}(\Theta_k,\theta) \tag{2}$$

where $Q_{k+1}(\Theta_k, \theta)$ is the Q-function for the $(k+1)$th utterance, and

$$L_{k+1}(\Theta_k,\theta) = \sum_j \sum_m P(\eta_{k+1} = j, \varepsilon_{k+1} = m \mid y_1^{k+1}, \Theta_k) \log p(y_{k+1} \mid j, m) \tag{3}$$

[0015]    Based on stochastic approximation, sequential updating equation is

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 Q_{k+1}^{(s)}(\Theta_k,\theta)}{\partial^2 \theta}\right]_{\theta=\theta_k}^{-1} \left[\frac{\partial l_{k+1}(\Theta_k,\theta)}{\partial \theta}\right]_{\theta=\theta_k} \tag{4}$$

[0016]    This says you get the newly estimated parameter $\theta_{k+1}$ based on $\theta_k$ minus the second derivative and the first derivative of the $Q$ function. $k$ here is the index of the utterance. This shows that at each utterance you can update the change following the channel or speaker change.

[0017]    We then apply this to the bias estimation to get sequential estimation of state-dependent biases. We introduce a state-dependent bias $l_j$ attached to each state j, we express the Gaussian power density function (pdf) of the state $j$ mixture $m$ as

$$b_{jm}(o_t) = N(o_t; \mu_{jm} + l_j, \sum_{jm}) \tag{5}$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}\left|\sum_{jm}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(o_t - \mu_{jm} - l_j)^T \sum_{jm}^{-1}(o_t - \mu_{jm} - l_j)}$$

[0018]    This equation 5 specifies the Gaussian distribution attached to the state $j$ and mixing component $m$. This equation shows at each state $j$ we have a bias $l_j$.

[0019]    Apply the block sequential estimation formula in equation 4,

$$l_j^{(k+1)} = l_j^{(k)} - \left[\frac{\partial^2 Q_{k+1}(\Theta_k, l_j)}{\partial^2 l_j}\right]_{l_j=l_j^{(k)}}^{-1} \left[\frac{\partial L_{k+1}(\Theta_k, l_j)}{\partial l_j}\right]_{l_j=l_j^{(k)}} \quad (6)$$

[0020]    Ignoring the items that are independent of $l_j$'s we define Q-function as

$$Q_{k+1}(\Theta_k, l_j) = \sum_{t=1}^{T^{k+1}} \sum_j \sum_m P(\eta_t = j, \varepsilon_t = m \mid o_1^{T^{k+1}}, \Theta_k) \log b_{jm}(o_t) \quad (7)$$

$$= \sum_{t=1}^{T^{k+1}} \sum_j \sum_m \gamma_{k+1,t}(j,m) \log b_{jm}(o_t) \quad (8)$$

where $\gamma_{k+1,t}(j,m) = P(\eta_t = j, \varepsilon_t = m \mid o_1^{T^{k+1}}, \Theta_k)$ is the probability that the system stays at

time t in state j mixture given the observation sequence $o_1^{T^{k+1}}$. This refers to the

probability P of being in state j, mixing component m given what we observe $O_1$ from 1

to $T^{k+1}$ and given old HMM $\Theta_k$.

[0021]    According to the definition,

$$\frac{\partial L_{k+1}(\Theta_k, l_j)}{\partial l_j} = \sum_m \sum_{t=1}^{T^{k+1}} \gamma_{k+1,t}(j,m) \sum_{jm}^{-1} (o_t - \mu_{jm} - l_j^{(k)}) \quad (9)$$

$$\frac{\partial^2 Q_{k+1}(\Theta_k, l_j)}{\partial^2 l_j} = -\sum_m \sum_{t=1}^{T^{k+1}} \gamma_{k+1,t}(j,m) \sum_{jm}^{-1} \quad (10)$$

[0022]    Therefore we arrive at the sequential updating relation for the state-

dependent biases in an utterance-by- utterance manner:

$$l_j^{(k+1)} = l_k^{(k)} + \left[\sum_m \sum_{t=1}^{T^{k+1}} \gamma_{k+1,t}(j,m) \sum_{jm}^{-1}\right]^{-1} \left[\sum_m \sum_{t=1}^{T^{k+1}} \gamma_{k+1,t}(j,m) \sum_{jm}^{-1} (o_t - \mu_{jm} - l_j^{(k)})\right] \quad (11)$$

[0023] In this above equation it shows at each state j we have a bias bias $l_j$. We therefore have as many biases as we have states. There could be as much as 3000 states. For some applications this is too high a number. In some applications, we teach herein to tie the biases into several classes $i$ in order to achieve more reliable and robust estimation.

[0024] In this case, a modification of equation 11 to sum up the accumulations inside each class.

$$l_i^{(k+1)} = l_i^{(k)} + \left[ \sum_{j \in class\ i} \sum_m \sum_{t=1}^{T^{k+1}} \gamma_{k+1,t}(j,m) \Sigma_{jm}^{-1} \right]^{-1} \left[ \sum_{j \in class\ i} \sum_m \sum_{t=1}^{T^{k+1}} \gamma_{k+1,t}(j,m) \Sigma_{jm}^{-1} (o_t - \mu_{jm} - l_i^{(k)}) \right]$$

$$(12)$$

[0025] As illustrated in Figure 4 we have all of the states in different frames tied to the same bias.

[0026] In summary, the state-dependent bias is updated at each utterance observation $k$. The update consists in an additive correction, composed of two factors. The first factor is based on an average variance, weighted by the probability of occupancy. The second one is based on the average of normalized difference between the observed vector and the model (original mean vector plus a bias, which has been adjusted with the utterances observed so far), weighted by the probability of occupancy.

[0027] Referring to Figure 3 there is illustrated the method according to one embodiment of the present invention to modify the mean vectors. The method includes introducing an adjustable bias in the distribution parameter of the signals. The detector 37 detects this parameter for every utterance. Every time a new observation of the signal is available updating the bias by calculating at calculator 35 a new updated bias by adding a correction term to the old bias. The correction term is calculated based on the information

of both the current model parameters and the incoming signals. The correction term is also calculated on the information from all signals provided to the recognizer and all incoming observed signals. Therefore, every time we update we don't forget the past and the previous updates are taken into account. The signals are speech signals. As discussed previously the new available data could be based on any length, in particular, could be frames, utterances or every fixed time period such as 10 minutes of speech signal. The correction term is the product of two items: the first item could be any sequences whose limit is zero, whose summation is infinity and whose square summation is not infinity. And the second term is a summation of quantities weighted by a probability, the quantities are based on the divergence model parameter and observed signal. The bias can be defined on each HMM state as in equation 11, or can be shared among different states or can be shared by groups of states or it can be shared by all the distribution of the recognizer by tying together as in equation 12.